Prüfungen/Psychometrie (Vorträge)

V1-612 (174)

Stabile Antwortmuster bei Script Concordance Test Fragen in der Schweizer Facharztprüfung Allgemeine Innere Medizin

Daniel Stricker[1], Felicitas-Maria Lahner[1], Raphael Bonvin[2], Christoph Berendonk[1]

[1]Bern, Schweiz
[2]Lausanne, Schweiz

Fragestellung: Mit dem Script Concordance Test (SCT) soll die Fähigkeit zu klinischem Denken (clinical reasoning) geprüft werden [1], [2]. Jedoch wird das Fragenformat u.a. kritisiert, weil die Messzuverlässigkeit schwierig zu überprüfen ist. Insbesondere fehlen Angaben zu test-retest Reliabilitäten [3]. Ziel der vorliegenden Studie ist es, die Stabilität der Antwortmuster auf SCT Fragen zu untersuchen.

Methoden: In zwei Facharztprüfungen für Allgemeine Innere Medizin in der Schweiz (Juni 2014, November 2014) wurden (neben 100 Multiple Choice Fragen) jeweils 20 SCT Fragen eingesetzt, zehn davon waren in beiden Prüfungen identisch. Insgesamt nahmen 591 Kandidaten an den Prüfungen Teil (Juni:287, November:304). Alle Kandidaten stammten aus derselben Weiterbildungskohorte und konnten frei zwischen beiden Terminen wählen. Die SCT Items wurden auf einer fünfstufigen Skala beantwortet und mit aggregate Scoring [4] basierend auf einem Expertenpanel (N=26) bewertet.

Wir verglichen die wiederholten SCT-Items bezüglich der Verteilung der Antworten der Kandidaten über die unterschiedlichen Antwortalternativen.

Ergebnisse: Die mittlere Leistung der beiden Prüfungsgruppen in den wiederholten SCT Items ist identisch (Juni: M=7.44, SD=0.975; November: M=7.45, SD=0.939). Hochgerechnet auf 100 Fragen beträgt das Cronbach-a des SCT Teils.77 (Juni) resp. 67 (Nov.). Der Vergleich der Antwortmuster der beiden Prüfungsgruppen war für alle 10 wiederholten SCT Fragen identisch.

Diskussion: Diese Resultate legen nahe, dass die Messzuverlässigkeit der wiederverwendeten SCT Fragen als hoch einzustufen ist. Auch wenn mit dieser Untersuchung längst nicht alle methodischen Probleme der SCT Fragen im summativen Einsatz geklärt werden [3] ist das Resultat dennoch bemerkenswert, weil beide Gruppen unabhängig waren und zwischen den beiden Erhebungen 5 Monate lagen.

Take home messages: Diese Studie zeigt, dass der sorgfältige Einsatz von SCT Fragen zu stabilen Resultaten führen kann.

Literatur

1. Charlin B, van der Vleuten C. Standardized Assessment of Reasoning in Contexts of Uncertainty: The Script Concordance Approach. Eval Health Prof. 2004;27(3):304-319. DOI: 10.1177/0163278704267043
2. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: from theory to practice: AMEE guide no. 75. Med Teach. 2013;35(3):184-193. DOI: 10.3109/0142159X.2013.760036
3. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47(12):1175-1183. DOI: 10.1111/medu.12283
4. Bland AC, Kreiter CD, Gordon JA. The psychometric properties of five scoring methods applied to the script concordance test. Acad Med. 2005;80(4):395-399. DOI: 10.1097/00001888-200504000-00019

Bitte zitieren als: Stricker D, Lahner FM, Bonvin R, Berendonk C. Stabile Antwortmuster bei Script Concordance Test Fragen in der Schweizer Facharztprüfung Allgemeine Innere Medizin. In: Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA). Bern, 14.-17.09.2016. Düsseldorf: German Medical Science GMS Publishing House; 2016. DocV1-612.
DOI: 10.3205/16gma174, URN: urn:nbn:de:0183-16gma1745
Frei verfügbar unter: http://www.egms.de/en/meetings/gma2016/16gma174.shtml

V1-643 (175)

Psychometrische Gütekriterien von Multiple-Choice-Examen in Abhängigkeit der Anzahl Kandidaten und Items: Ab welchen Stichprobengrössen sind die Gütekriterien vertrauenswürdig?

Rainer Hofer, Sören Huwendiek

Bern, Schweiz

Fragestellung: In Prüfungsanalysen mit einer kleinen Anzahl von Kandidaten und/oder Items wird die Aussagekraft der psychometrischen Gütekriterien in der Itemanalyse respektive in der Interpretation zum Teil nicht ausreichend mitberücksichtigt. In der vorliegenden Studie wurde der Frage nachgegangen, ab welcher Anzahl Kandidaten und Items die Gütekriterien ausschliesslich in dem als vertrauenswürdig bestimmten Intervall liegen.

Methode: Der Studie lagen die Daten der Eidgenössischen Prüfung Humanmedizin des Jahres 2014 zugrunde, bei der 592 deutschsprachige Kandidaten 300 Multiple-Choice-Fragen beantworteten. Die Daten der 269 französischsprachigen Kandidaten wurden nicht berücksichtigt, um soziokulturelle Einflüsse bestmöglich auszuschliessen. Als Ausgangslage dienten die Werte der Gütekriterien (wie Reliabilität, Schwierigkeit, Standardmessfehler, Trennschärfe) über alle 592 Kandidaten und alle 300 Items. Für diese Werte wurden die 95%-Vertrauensintervalle bestimmt.

Mittels Bootstrapping [1] wurden danach 100 Stichproben aus der Grundgesamtheit gezogen. Über die 100 Ziehungen wurden die Gütekriterien gemittelt, deren Vertrauensintervalle berechnet und diese mit den Ausgangswerten verglichen.

Das Bootstrapping wurde iterativ mit jeweils um 10 Kandidaten und/oder 10 Items reduzierten Stichproben wiederholt. In der entgegengesetzten Iteration wurden mit Stichproben von 10 Kandidaten und 10 Items gestartet und anschliessend die Stichproben nach dem Zufallsprinzip jeweils um 10 Kandidaten und/oder 10 Items vergrössert. Die Analyse wurde mit dem Statistikpaket R durchgeführt.

Ergebnisse: Die Daten werden bis zum Beitrag zu Ende ausgewertet sein. Es wird aufgezeigt werden, ab welchen Stichprobengrössen (Kandidaten, Items) die Gütekriterien „vertrauenswürdig" sind.

Diskussion: Die Ergebnisse werden anhand der Literatur diskutiert werden.

Take-Home-Messages: Entsprechend der Ergebnisse werden Take Home Messages formuliert werden.

Literatur

1. Efron B. Bootstrap methods: another look at the jackknife. Ann Statist. 1979;7:1–26. DOI: 10.1214/aos/1176344552

Bitte zitieren als: Hofer R, Huwendiek S. Psychometrische Gütekriterien von Multiple-Choice-Examen in Abhängigkeit der Anzahl Kandidaten und Items: Ab welchen Stichprobengrössen sind die Gütekriterien vertrauenswürdig? In: Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA). Bern, 14.-17.09.2016. Düsseldorf: German Medical Science GMS Publishing House; 2016. DocV1-643. DOI: 10.3205/16gma175, URN: urn:nbn:de:0183-16gma1757
Frei verfügbar unter: http://www.egms.de/en/meetings/gma2016/16gma175.shtml

V1-662 (176)

Einfluss von unterschiedlichen Bewertungs-Algorithmen für Kprim Fragen auf psychometrische Charakteristiken von Prüfungen

Felicitas-Maria Lahner[1], Zineb Nouns[1], Martin R. Fischer[2], Sören Huwendiek[1]

[1]Bern, Schweiz
[2]München, Deutschland

Fragestellung/Zielsetzung: Die Vor- und Nachteile unterschiedlicher Bewertungsalgorithmen von Kprim-Fragen [1], [2], ist nicht eindeutig geklärt. Diese Studie untersucht den Einfluss verschiedener Bewertungs-Algorithmen für Kprim-Fragen auf deren psychometrische Parameter und vergleicht diese mit denen von Typ A-Fragen.

Methoden: Wir untersuchten an einer Stichprobe von 38 Prüfungen (998 Kprim und 2163 Typ A Items, durchschnittlich 225 Kandidaten/Prüfung) zweier Schweizer Fakultäten sowie der Eidgenössischen Prüfung den Einfluss unterschiedlicher Bewertungs-Algorithmen für Kprim Fragen auf Reliabilität, Trennschärfe, Schwierigkeit und die Gesamtpunktzahl.

Wir verglichen drei Bewertungs-Algorithmen für Kprim Items mit 4 Antwortmöglichkeiten:

1. Viertelpunkt-Bewertung (VP): für jede richtige Teilantwort ¼ Punkt
2. Halbpunkt-Bewertung (HP): ½ Punkt wenn mehr als die Hälfte, 1 Punkt, wenn alle Teilantworten richtig beantwortet.
3. Ganzpunkt-Bewertung (GP): 1 Punkt wenn alle Teilantworten richtig beantwortet.

Zum Vergleich wurden Typ A Fragen miteinbezogen. Die Bewertungs-Algorithmen verglichen wir mit Varianzanalysen für wiederholte Messungen bzw. Friedmann Tests falls die Voraussetzungen für Varianzanalysen nicht erfüllt wurden.

Ergebnisse: VP und HP führen zu signifikant höheren Reliabilitäten und Trennschärfen im Vergleich zu GP und Typ A. Im Bezug auf die Itemschwierigkeit unterscheiden sich alle Bewertungs-Algorithmen signifikant, wobei VP leichteren und GP zu den schwierigeren Items führt. HP führt zu leichteren Items als Typ A. Bei der Gesamtpunktzahl zeigt sich, dass Kandidaten mit VP im Durchschnitt signifikant mehr Punkte erreichen als mit HP oder GP.

Diskussion: Bewertungs-Algorithmen mit Teilpunkten führen zu besseren psychometrischen Charakteristiken. Dies zeigt sich auch in anderen Studien zu Fragen mit Mehrfachantworten wie z.B. Pick-N [3] oder kleineren Studien mit Kprim-Fragen [4].

Take home message: Bewertungen mit Teilpunkten führen bei Kprim-Fragen zu besseren psychometrischen Charakteristiken.

Literatur

1. Javid L. The Comparison between Multiple-choice (MC) and Multiple True-false (MTF) Test Formats in Iranian Intermediate EFL Learners' Vocabulary Learning. Procedia Soc Behav Sci. 2014;98:784-788. DOI: 10.1016/j.sbspro.2014.03.482
2. Mobalegh A, Barati H. Multiple True-false (MTF) and Multiple-choice (MC) Test Formats: A Comparison between Two Versions of the Same Test Paper of Iranian NUEE. J Lang Teach Res. 2012;3(5):1027-1037. DOI: 10.4304/jltr.3.5.1027-1037
3. Bauer D, Holzer M, Kopp V, Fischer MR. Pick-N multiple choice-exams: a comparison of scoring algorithms. Adv Health Sci Educ Theory Pract. 2011;16(2):211-221. DOI: 10.1007/s10459-010-9256-1
4. Albanese MA, Sabers DL. Multiple True-False Items: A Study of Interitem Correlations, Scoring Alternatives, and Reliability Estimation. J Educ Meas. 1988;25(2):111-123. DOI: 10.1111/j.1745-3984.1988.tb00296.x

Bitte zitieren als: Lahner FM, Nouns Z, Fischer MR, Huwendiek S. Einfluss von unterschiedlichen Bewertungs-Algorithmen für Kprim Fragen auf psychometrische Charakteristiken von Prüfungen. In: Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA). Bern, 14.-17.09.2016. Düsseldorf: German Medical Science GMS Publishing House; 2016. DocV1-662.
DOI: 10.3205/16gma176, URN: urn:nbn:de:0183-16gma1768
Frei verfügbar unter: http://www.egms.de/en/meetings/gma2016/16gma176.shtml